A Mobile Robot Visual Perception System based on Deep Learning Approach

Aleksandar Jokić, Lazar Đokić, Milica Petrović and Zoran Miljković

***Abstract*—In this paper, we present the novel mobile robot perception system based on a deep learning framework. The hardware subsystem consists of an Nvidia Jetson Nano development board integrated with two parallelly positioned Basler daA1600-60uc cameras, while the software subsystem is based on the convolutional neural networks utilized for semantic segmentation of the environment scene. A Fully Convolutional neural Network (FCN) based on the ResNet18 backbone architecture is utilized to provide accurate information about machine tool models and background position in the image. FCN model is trained on our custom-developed dataset of a laboratory model of manufacturing environment and implemented on mobile robot RAICO (Robot with Artificial Intelligence based COgnition).**

***Index Terms*—Deep learning; Perception System; Mobile robot; Semantic Segmentation.**

I. Introduction

Modern mobile robot sensors (e.g., cameras or lidars) provide a rich amount of data about the current state of the environment. However, the way the data is interpreted and transferred into useful information has been an active area of research in the last two decades. Deep learning, or more precisely, Convolutional Neural Networks (CNNs), represent one of the most promising methodologies that can enable mobile robots to understand and interact with their environment in a more sophisticated manner [1]. The main disadvantage of CNNs is the requirement for a substantial amount of computation power for real-time implementation. Fortunately, several modern single-board computers or hardware accelerators provide enough computing power to deploy low-weight CNNs for real-time implementation.

The authors [2] developed the visual perception system utilized for navigation in the indoor environment based on CNNs for scene classification. They deployed shallow CNN to achieve real-time mobile robot navigation within the environment in both dynamic and static conditions. The development of the CNN model capable of determining the 3D physical properties of objects in the scene is presented in [3]. The authors propose using the learned properties to predict the outcome of the dynamic events in the environment. This type of perception system can be beneficial for mobile robots employed in highly dynamic environments. The authors of [4] developed a complex cleaning mobile robot perception system with two submodules, one based on Bayesian filtering of data from 2D lidar, 3D lidar, and RGB-D camera used for human detection and tracking, and the second one for obstacle and dirt detection based on two RGB-D cameras and 3D lidar sensor. In [5], the authors developed a perception system based on a monocular camera for data acquisition and SURF point feature extraction method integrated with neural extended Kalman filter for simultaneous localization and mapping of mobile robot's position and orientation. Another promising methodology for developing perception systems for mobile robots is a cooperative perception [6], where the perception of one autonomous agent depends on the perception of other nearby agents. Camera information obtained by one mobile robot (agent) can be propagated to the others if their relative pose is estimated well. Moreover, the authors implemented the camera models that share information regarding pedestrian detection. The mobile robot developed with the aim to be applied for manufacturing purposes was proposed in [7]. Safe indoor navigation is provided by the perception system based on a 2D camera, 3D camera, laser scanner, ultrasonic sensors, and internal measurement unit. On the one side, safe navigation is provided by a laser scanner and ultrasonic sensors, while a 3D camera is used for the correction step in the Kalman filter state estimation. Moreover, the 2D camera is utilized for adaptive path planning by detecting lines on the ground.

This work presents the novel semantic segmentation-based mobile robot perception system implemented on our own developed mobile robot RAICO. The modified ResNet18 backbone architecture is integrated with RAICO's sensory subsystem, which contains two parallelly positioned Basler daA1600-60uc cameras. The proposed perception system provides RAICO with the ability to safely navigate the laboratory model of the manufacturing environment by knowing the position of machine tools in the image plane.

Aleksandar Jokić, PhD student, University of Belgrade - Faculty of Mechanical Engineering, Department of Production Engineering, Laboratory for industrial robotics and artificial intelligence (ROBOTICS&AI), Kraljice Marije 16, 11120 Belgrade 35, The Republic of Serbia (ajokic@mas.bg.ac.rs).

Lazar Đokić, PhD student, University of Belgrade - Faculty of Mechanical Engineering, Department of Production Engineering, Laboratory for industrial robotics and artificial intelligence (ROBOTICS&AI), Kraljice Marije 16, 11120 Belgrade 35, The Republic of Serbia (ldjokic@mas.bg.ac.rs).

Dr. Milica Petrović, Assistant Professor, University of Belgrade - Faculty of Mechanical Engineering, Department of Production Engineering, Laboratory for industrial robotics and artificial intelligence (ROBOTICS&AI), Kraljice Marije 16, 11120 Belgrade 35, The Republic of Serbia (mmpetrovic@mas.bg.ac.rs).

Dr. Zoran Miljković, Full Professor, University of Belgrade - Faculty of Mechanical Engineering, Department of Production Engineering, Laboratory for industrial robotics and artificial intelligence (ROBOTICS&AI), Kraljice Marije 16, 11120 Belgrade 35, The Republic of Serbia (zmiljkovic@mas.bg.ac.rs).

The paper is structured as follows. Section two is devoted to a thorough explanation of the considered CNN model and its training procedure. The third Section describes the experimental results and perception system evaluation while concluding remarks are presented in the fourth Section.

II. Mobile Robot Perception System

The mobile robot perception system is developed with two parallelly mounted Basler daA1600-60uc cameras facing downwards with an inclination angle of 30° and a baseline of 12.5 cm. The combination of mentioned cameras with *Evetar M118B0418W* lenses (4 mm focal length) provides a large angle-of-view of the scene, approximately W×H=84°×68°. The cameras are connected via USB3.0 to the Jetson Nano development board for image acquisition. The whole perception system is positioned on the top of the mobile robot RAICO (Fig. 1).

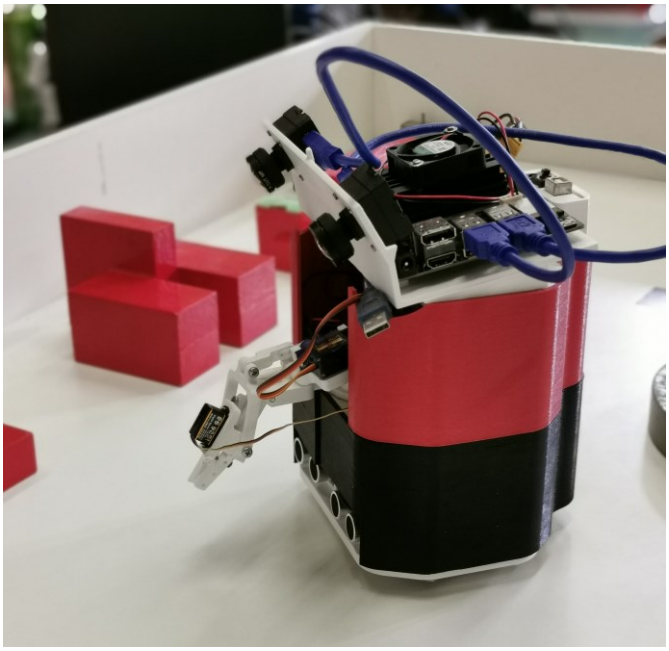

Fig. 1. Mobile robot RAICO with assembled perception system

The main component of the perception system is the Fully Convolutional neural Network (FCN) used for semantic segmentation. To enable the processing in near real-time, the selected backbone is ResNet18 based network. The network is trained by utilizing our custom-developed dataset for semantic segmentation. Image data is acquired within the laboratory model of a manufacturing environment, while the segmentation masks are hand-labeled. The dataset consists of densely labeled images with four machine tool classes and the background class. The dataset contains 125 images divided for training and testing in the 80/20 ratio. The sample of the dataset is shown in Fig. 2.

Hard data augmentation is carried out on the images used for training to improve neural network generalization. Mobile robot RAICO uses real-world noise-prone cameras that can significantly impact the accuracy of neural networks [8]. Therefore, the first augmentation is performed with Gaussian noise added to the images. Two Gaussian noise levels have been introduced for all the images used for training. The first level contains the noise with zero mean and variance in the range of 0.002-0.004, and the second one has a variance of 0.002-0.011 (sample of image with Gaussian noise is shown in Fig. 3).

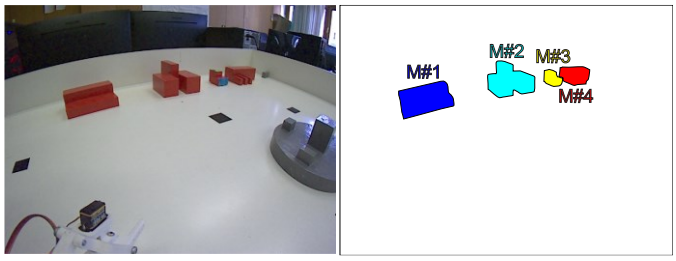

Fig. 2. Sample of the custom-developed dataset

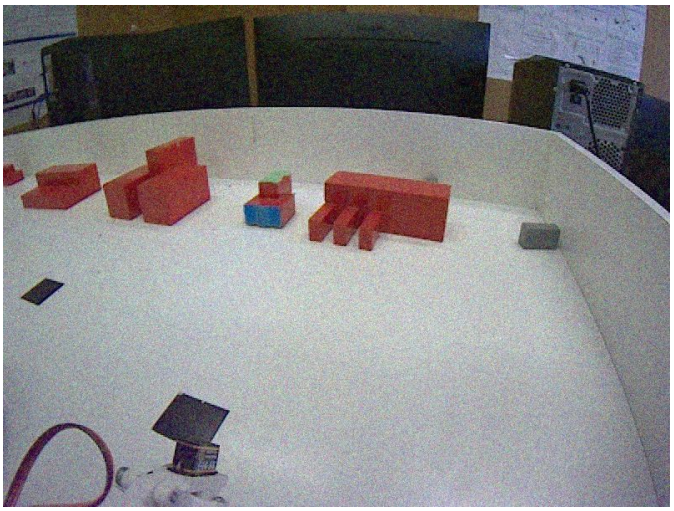

Fig. 3. Image with Gaussian noise

Besides the Gaussian noise procedures, we also applied three data augmentation procedures: (i) horizontal flips (10% of the images), (ii) random crops with a scale of 0.7 (10% of the images), and (iii) the complete image pixel intensities change in the range of 0.8-1.2 (10% of images); this procedure results in images with different illumination intensity levels, which can realistically occur during the different parts of the day. After data augmentation is done, the considered dataset contains 370 images used for training neural networks.

The details about utilized architecture are presented in Fig. 4. Different blocks of layers are presented with rectangles of different colors, while the parameters for those layers are presented within the rectangle. **W** represents the weight matrix dimensions, S is the stride value, and P represents the padding value. Convolution, BatchNormalization, and ReLU layers are presented with blue blocks. The green block presents the MaxPooling layer, while the convolution and BatchNormalization block is presented with the brown

rectangle. Finally, the adding layer in combination with the ReLU activation layer is presented with orange. Input images have 800×600×3 resolution, while the output semantic mask has the dimension of 19×25. The probabilities of the class prediction are calculated by utilizing the Softmax activation function (1), while the utilized loss function is Cross-entropy (2).

$$s_i = \frac{e^{y_i}}{\sum_{i=1}^{N} e^{y_i}} \tag{1}$$

$$\ell(\mathbf{s},\mathbf{c}) = -\sum_{i}^{N} c_i \log(s_i) \tag{2}$$

Where **y** represents the output vector of the neural network, i is the current element of the output vector, N is a total number of classes (and the number of elements in the output vector), s_i is the output of the softmax function for each element, **c** represents one-hot vector for the correct class of the current input vector, and ℓ represents the loss function value.

The training is carried out by PyTorch v1.6.0 with Stochastic gradient descend and the momentum of 0.9. The initial learning rate is η=0.01 with the changing schedule defined with (3):

$$\eta^{\text{new}} = \eta^{\text{old}} \cdot \left(1 - \frac{\text{current\_epoch}}{\text{max\_epoch}}\right)^{0.9}. \tag{3}$$

It is important to note that current_epoch is enumerated from 0 to (max_epoch – 1). For the experimental research presented in this paper, the maximum number of epochs is 30, while the mini-batch size is 4. Lastly, the regularization technique is utilized with a weight decay of 0.0001. Training is performed on Nvidia RTX 1660 GPU with 6GB of RAM.

Since the Nvidia Jetson nano is an edge device with limited processing power (NVIDIA 128-core Maxwell GPU), the whole FCN network with encoder and decoder parts could not be implemented in real-time. Therefore, the authors propose to maintain the output of the backbone network and directly calculate the semantic mask with the output resolution instead of deconvolving that information to acquire prediction with the same resolution as input. Having that in mind, the output mask is considerably smaller in resolution than an input image. However, the achieved accuracy is entirely satisfactory.

III. Experimental Results

We have trained the FCN-ResNet18 model on our custom-developed dataset for semantic segmentation of laboratory model of a manufacturing environment. Moreover, the trained model is implemented in the perception system on the mobile robot RAICO. Two metrics utilized to analyze the generalization performance of the FCN-ResNet18 model are Global accuracy and Intersection over Union (IoU). The training results for each class, as well as for the whole dataset, are presented in Table I.

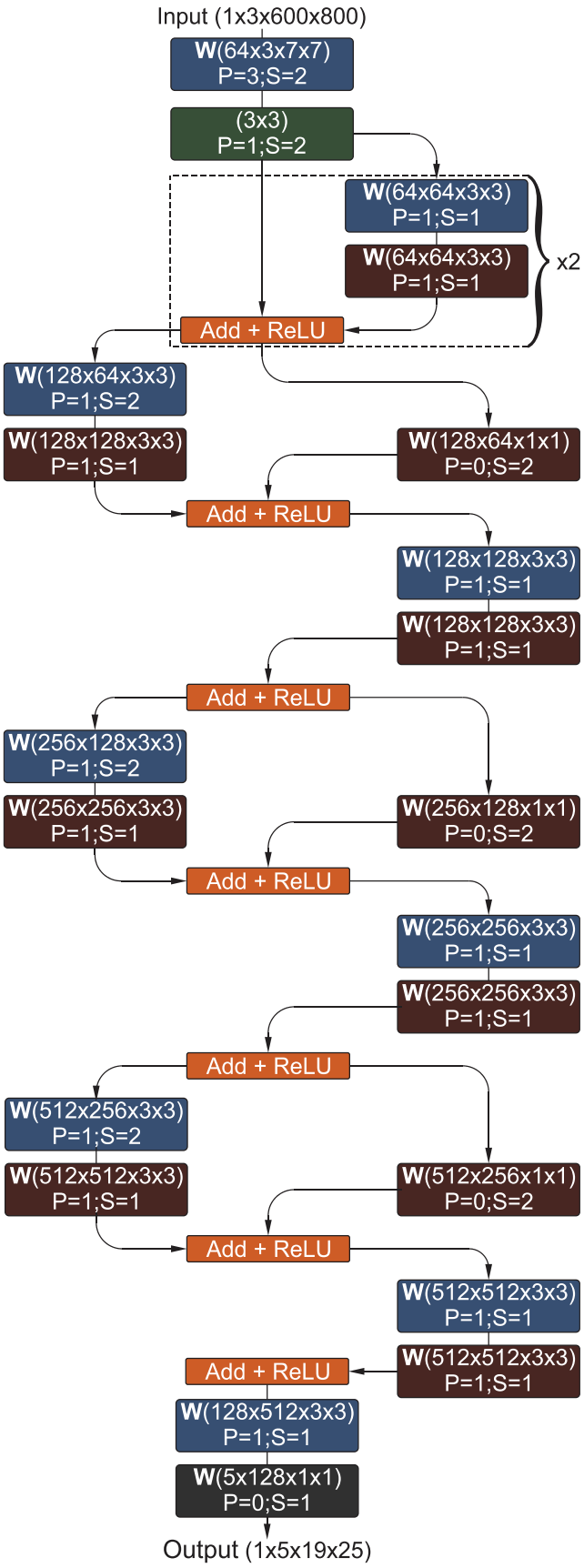

Fig. 4. Architecture of the FCN-ResNet18 model

TABLE I
THE EXPERIMENTAL RESULTS OF THE SEMANTIC SEGMENTATION MODEL

Accuracy measures	Background	M#1	M#2	M#3	M#4
Global per-class accuracy [%]	96.8	68.9	88.6	91.4	94.4
Per-class IoU	96.1	58.2	69.4	45.9	58.6
Mean global accuracy =96.0			Mean IoU = 65.6		

As shown in Fig. 5, there is a significant class imbalance in the considered dataset, as in most semantic segmentation datasets. The dominant class in the images is the background.

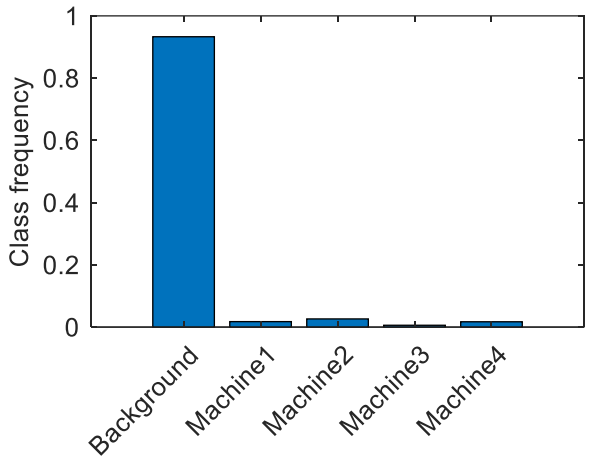

Fig. 5. Class frequency in the custom dataset

Having that in mind, the highest accuracy is achieved for the class with most samples, even though the authors have added the class weights that are inversely proportional to the class frequencies. Moreover, the worst results (for IoU metric) are achieved for Machine3 (M#3) since it is the smallest machine and therefore occupies the smallest percentage of the scene. Interestingly, global accuracy for M#1 is the smallest compared to all the other classes. The authors further investigated this occurrence and presented the overlay view of two test images and their semantic masks generated by the FCN network (Fig. 6). As it can be seen, the network misclassified half of the M#1 in the first image in Fig. 6. Furthermore, in the image with occlusions, part of the M#1 is misclassified and labeled as M#4.

Achieved mean global accuracy is 96.0%, which is a promising result; however, the mean IoU measure of 65.6 is much more representative of the actual generalization capabilities of the FCN model.

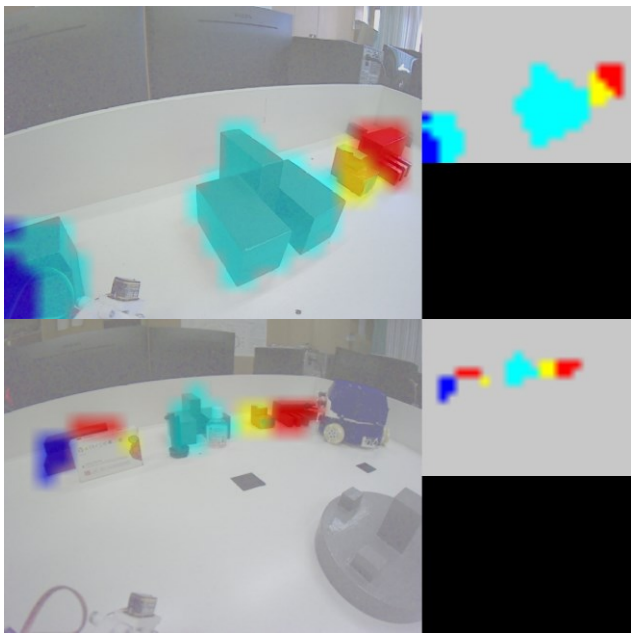

Fig. 6. Test images overlayed with semantic maps

To further test the accuracy of the trained network, the model is implemented on mobile robot RAICO and tested online by the real-time acquisition of images and their semantic segmentation. Fig. 7 presents few images acquired and segmented by the FCN model in a real-world scenario. To increase the effectiveness of the FCN model, it is transformed to an ONNX format and optimized by utilizing Nvidia TensorRT.

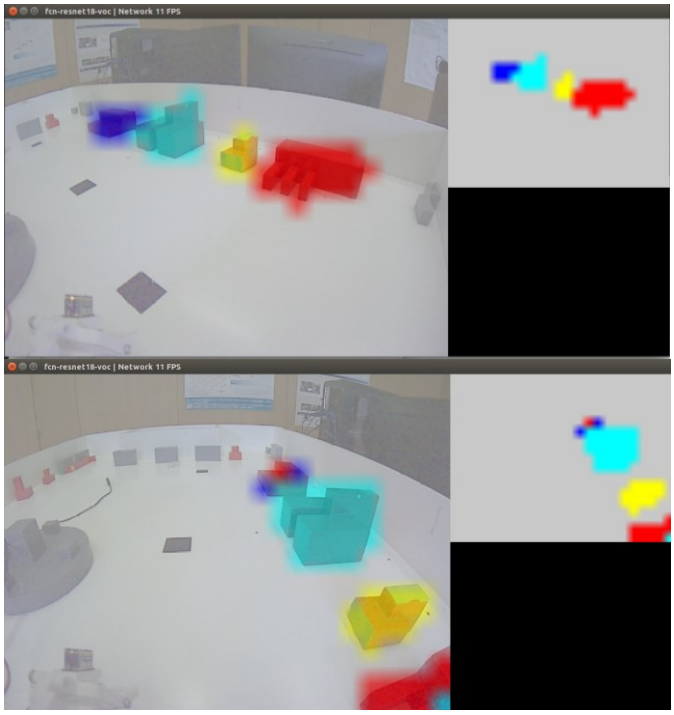

Fig. 7. Testing of implemented FCN model

From Fig. 7, it can be seen that the considered FCN model achieves acceptable accurate segmentation results, with minor errors on machines that are either far away, occluded by other machines, or only partially visible. Furthermore, the model is implemented with 11FPS, which is acceptable for a mobile robot with low-velocity profiles.

IV. Conclusion

This paper proposes the new perception system of mobile robot RAICO based on a Fully Convolutional neural Network with ResNet18 backbone architecture. Training of the neural network model is carried out on a custom-developed dataset for semantic segmentation of the laboratory model of the manufacturing environment. The perception system is integrated with the Nvidia Jetson Nano development board and two Basler dart cameras and configured as a standalone-edge device. After the training procedure is completed, the model is implemented on the mobile robot RAICO, with the achieved accuracy measures of 65.6 for mean IoU and 96.0 for the global accuracy. The implemented system works in a near real-time manner achieving approximately 11FPS. Future research directions could include creating a larger dataset with more classes of manufacturing entities, as well as developing a novel, faster architecture for semantic segmentation capable of running real-time on Jetson nano.

Acknowledgment

This work has been financially supported by the Ministry of Education, Science and Technological Development through the project “Integrated research in macro, micro, and nano mechanical engineering – Deep learning of intelligent manufacturing systems in production engineering” (contract No. 451-03-9/2021-14/200105), and by the Science Fund of the Republic of Serbia, grant No. 6523109, AI – MISSION 4.0, 2020 – 2022.

References

[1] J. Shabbir and T. Anwer, “A survey of deep learning techniques for mobile robot applications,” *arXiv Prepr. arXiv1803.07608*, 2018.

[2] T. Ran, L. Yuan, and J. B. Zhang, “Scene perception based visual navigation of mobile robot in indoor environment,” *ISA Trans.*, vol. 109, pp. 389–400, 2021.

[3] J. Wu, I. Yildirim, J. J. Lim, B. Freeman, and J. Tenenbaum, “Galileo: Perceiving physical object properties by integrating a physics engine with deep learning,” Proc. Advances in neural information processing systems (NIPS 2015), vol. 28, pp. 127–135, 2015.

[4] Z. Yan, S. Schreiberhuber, G. Halmetschlager, T. Duckett, M. Vincze, and N. Bellotto, “Robot perception of static and dynamic objects with an autonomous floor scrubber,” *Intell. Serv. Robot.*, vol. 13, pp. 403–417, 2020.

[5] Z. Miljković, N. Vuković, M. Mitić, and B. Babić, “New hybrid vision-based control approach for automated guided vehicles,” *Int. J. Adv. Manuf. Technol.*, vol. 66, no. 1–4, pp. 231–249, 2013.

[6] S. Sridhar and A. Eskandarian, “Cooperative perception in autonomous ground vehicles using a mobile-robot testbed,” *IET Intell. Transp. Syst.*, vol. 13, no. 10, pp. 1545–1556, 2019.

[7] J. Qian, B. Zi, D. Wang, Y. Ma, and D. Zhang, “The design and development of an omni-directional mobile robot oriented to an intelligent manufacturing system,” *Sensors*, vol. 17, no. 9, Article Number: 2073, 2017.

[8] D. M. Chan and L. D. Riek, “Object proposal algorithms in the wild: Are they generalizable to robot perception?,” Proc. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2601–2607, 2019.